Med Biol Eng Comput (2016) 54:1831–1844
DOI 10.1007/s11517-016-1477-x

ORIGINAL ARTICLE

# Prediction of intrinsically disordered regions in proteins using signal processing methods: application to heat-shock proteins

**Vuk Vojisavljevic**[1] · **Elena Pirogova**[1]

Received: 20 August 2015 / Accepted: 25 February 2016 / Published online: 1 April 2016


**Abstract** Heat-shock protein (HSP)-based immunotherapy is believed to be a promising area of development for cancer treatment as such therapy is characterized by a unique approach to every tumour. It was shown that by inhibition of HSPs it is possible to induce apoptotic cell death in cancer cells. Interestingly, there are a great number of disordered regions in proteins associated with cancer, cardiovascular and neurodegenerative diseases, signalling, and diabetes. HSPs and some specific enzymes were shown to have these disordered regions in their primary structures. The experimental studies of HSPs confirmed that their intrinsically disordered (ID) regions are of functional importance. These ID regions play crucial roles in regulating the specificity of interactions between dimer complexes and their interacting partners. Because HSPs are overexpressed in cancer, predicting the locations of ID regions and binding sites in these proteins will be important for developing novel cancer therapeutics. In our previous studies, signal processing methods have been successfully used for protein structure–function analysis (i.e. for determining functionally important amino acids and the locations of protein active sites). In this paper, we present and discuss a novel approach for predicting the locations of ID regions in the selected cancer-related HSPs.

**Keywords** Intrinsic disorder · Cancer · HSPs · Signal processing · Active/binding site

## 1 Introduction

Proteins can express their biological activity by achieving a special stable three-dimensional (3D) structure. This paradigm, however, has been challenged recently and appears to be not entirely accurate for all proteins. In the last decades, a large number of individual proteins, having no fixed 3D structures but still able to express important biological functions, have been discovered. These proteins were called intrinsically disordered (ID) proteins. There are stable and ordered as well as flexible regions within the structures of ID proteins. These ID regions can vary significantly in size. As reported, more than one-third of eukaryotic proteins have been shown to contain ID regions of over 30 amino acids in length in their structures. This structural flexibility, as was shown in experimental studies, leads to a major functional advantage for ID proteins [15, 18, 31].

Research studies demonstrated that proteins with ID regions in their structures are involved in the biological processes such as cell-cycle control, regulation, recognition, and signalling [22, 33, 43, 45, 46]. Due to existence of the "unlocked" regions within their structures, which present flexible large surface areas for interactions, the ID proteins can bind to a broad range of ligands (other proteins, small molecules, membranes, and nucleic acids) [3, 16, 24, 36]. Moreover, they can simultaneously interact with multiple proteins. Since interactions are controlled by protein concentration, dynamic interactions of ID proteins with multiple partners result in significant changes to their concentrations. The multiple interactions of ID proteins with their targets are characterized by relatively high specificity

**Electronic supplementary material** The online version of this article (doi:10.1007/s11517-016-1477-x) contains supplementary material, which is available to authorized users.

✉ Elena Pirogova
elena.pirogova@rmit.edu.au

Vuk Vojisavljevic
vuk.vojisavljevic@rmit.edu.au

[1] Biomedical Engineering, School of Engineering, RMIT University, Melbourne, VIC 3001, Australia

and low affinity. This binding mechanism allows ID proteins to rapidly initiate a signalling process. The disordered regions enable ID proteins to mediate specific recognition of their interactive partners and also regulate the interaction in space and time [34]. Importantly, ID proteins, upon binding to a target molecule, become structured and stable [16].

The interactive capabilities of ID proteins have attracted much scientific interest, and thus, their characterization has become one of the fastest growing areas of protein science. Due to ID proteins being overexpressed in major disease pathways, they present desirable targets for inhibition. Understanding protein–protein interactions and possibilities of their inhibition has been improved in recent years [15, 31]. It was shown that the energy of protein–protein interaction is not evenly distributed over a large contact area but rather is focused in regions, whose areas are better accessible for contact/binding by a small molecule [43]. Therefore, overexpression of ID proteins in particular diseases, i.e. cancer, can be inhibited by targeted interactions of specific small molecules that can bind with high specificity to the disordered regions within the ID proteins.

There is a large number of computational software tools developed for predictions of disordered regions in ID proteins. Over 60 methods for computational prediction of protein disorder from sequence have been made publicly available. These prediction methods aim to identify disordered regions in ID proteins through analysis of amino acid sequences using mainly the physico-chemical properties of the amino acids, sequence complexity, amino acid composition or evolutionary conservation [12, 14, 37, 38]. Disorder prediction algorithms take into account the characteristic features of unstructured proteins and have been shown to be successful (almost 80 % accuracy), especially in the case of large regions [38]. The available prediction algorithms are based on different approaches: analysis of primary sequence composition; neural networks trained on X-ray structure data; local amino acid composition, flexibility, and hydropathy; neural networks trained on NMR solution-based data; cascaded support vector machine classifiers trained on Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) profiles, etc. According to the results of CASP11 (11th Community-Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction), the best prediction groups successfully identified 50–70 % of the disordered residues with false-positive rates from 3 to 16 % [23].

Heat-shock proteins (HSPs) are large proteins present in all living cells and play important functions in cell-cycle control and signalling. They also protect cells against stress (apoptosis). For example, HSPs are activated when a cell undergoes an environmental stress, such as oxygen deprivation, or is affected by temperature change (heat or cold stress). If there are no stress conditions, HSPs function as chaperones by preventing protein's misfolding or refold distorted or denaturated proteins, and help new proteins to fold into a shape required for their biological activity. HSPs also reshuffle proteins within a cell and transport old proteins to "garbage disposals" inside the cell. Extracellular HSPs are also important in immune defence as they recognize diseased cells and alert the immune system. Twenty-five years ago HSPs were identified as the key elements responsible for protecting animals from cancer, with anti-tumour vaccine development studies continuing today. It is known that HSPs are overexpressed in a wide range of human cancers. They are involved in tumour cell proliferation, differentiation, invasion, metastasis, and death. HSP-based immunotherapy is believed to be one of the most promising areas of developing cancer treatment technology that is characterized by a unique approach to every tumour [4–8, 19].

Considering their functional importance, this study is aimed to determine computationally the active/binding sites and ID regions in cancer-related HSP family (HSP27, HSP60, HSP70, and HSP90 proteins) by employing signal processing analysis methods. Our computational predictions for these selected HSP sequences are compared with the experimental data and predictions obtained using the selected predictors, MobiDB [13] and PONDR [32], the gold-standard tools in computational analysis of protein disorder. PONDR tool presents a series of neural network predictors (NNPs) that use amino acid sequence data to predict disorder in a given region. These neural networks use sequence attributes (fractional composition of particular amino acids, hydropathy, or sequence complexity) taken over windows of 9–21 amino acids. The attributes are averaged over these windows, and then the averaged values are used to train the neural network during predictor construction; the same values are used as inputs to make predictions. The NNPs are trained on selected sets of ordered and disordered protein sequences to enable them to generalize to new sequences. NNP output values are then smoothed over a sliding window of 9 amino acids. If an amino acid value exceeds or matches a certain threshold, then this residue is considered disordered. Another predictor selected for this study, MobiDB, presents a centralized resource for annotations of intrinsic protein disorder and includes three levels of annotation: *manually curated, indirect and predicted*. By combining all three levels into a consensus annotation, MobiDB tool provides the best possible picture of the "disorder landscape" of a given protein of interest. The recently updated MobiDB 2.0 tool [30] uses three biophysical predictors (IUPred-short, IUPred-long, and GlobPlot) and seven machine learning predictors (DisEMBL-465, DisEMBL-HL, Espritz-DisProt, Espritz-NMR, Espritz-xray, JRONN and VSL2b). Consensus prediction is formed by applying a

majority vote on the 10 predictors, when there is no high-quality information available from NMR, X-ray or DisProt data. Some of the predictors used in the MobiDB tool are [30]:

- IUPred is based on analysis of the pairwise energy content estimated from amino acid composition,
- DisEMBL and GlobPlot are based on alignment with known ID sequences,
- ESpritz is based on bidirectional recursive neural networks that are trained on three different flavours of disorder (including a novel NMR flexibility predictor),
- VSL2b and JRONN, also based on neural networks, can calculate consensus in the results of different algorithms applied to extract the most probably disordered amino acids regions.

In this study, the predictive capabilities of the proposed RRM-SPWVD approach are compared with the predictions generated by MobiDB2.0 and PONDR using the selected cancer-related HSPs as protein examples.

## 2 Materials and methods

### 2.1 Resonant recognition model

Signal processing methods have been used successfully in analysis of biological signals (ECG, EEG, EMG, proteins, etc.) [11, 17]. Of particular importance to this study is the possibility of applying signal processing techniques to analysis of macromolecules, i.e. proteins and DNA. The resonant recognition model (RRM) [9] belongs to the computational methods that aim at elucidating protein biological function from analysis of its protein primary structure—a polypeptide chain of amino acids. The RRM is a physico-mathematical approach designed for analysis of structure–function relationships in different proteins and their mutual interactions (theory can be applied to analysis of DNA and RNA). It is generally understood that protein's biological function can be described as its ability to bind to a specific ligand. This selective interaction of a protein with its target presents a multistage process that includes specific biorecognition, chemical binding, and energy transfer. The RRM [9, 25] concepts state that the selectivity of protein interaction is defined within a protein's primary structure, an amino acid sequence. According to the RRM principles, by analysing a protein primary structure, the critical information about its functionality can be obtained. The RRM postulates that protein (DNA) interactions present a resonant energy transfer between the interacting molecules at the frequency specific for each observed function/interaction [10, 26–29].

The model employs signal processing methods to analyse protein primary structures. Firstly, the original protein primary sequence is converted into a numerical sequence by assigning to each amino acid in the sequence a physical parameter value relevant to the protein's biological activity. Electron ion interaction potential (EIIP) is used as a physical property. The EIIP parameter describes the average energy states of all valence electrons in a particular amino acid, and its values for each amino acid were calculated from the general model of pseudopotentials [39]. The resultant numerical series is then analysed using digital signal processing, Fourier transform, which transform the signal into a single spectrum. To determine the common frequency components in the spectra for a group of proteins, the multiple cross-spectral function is used. Peaks in this function denote common frequency components for the protein sequences analysed (block diagram of the RRM approach is shown in Appendix IV, Supplementary Materials). The normalized intensity of the prominent peaks represents a signal-to-noise ration and depends on a length and a number of proteins used in the multiple cross-spectral function. The RRM postulates that there is a significant correlation between spectra of the numerical presentation of amino acids and their biological activity [9, 25]. Through extensive computational studies utilizing the RRM theory, it was found that the RRM frequencies present the characteristic features of different protein biological functions or interactions [9, 10, 25–29]. These characteristic RRM frequencies were shown to be relevant parameters for mutual recognition between biomolecules and are important in describing the selectivity of interaction between proteins and their substrates or targets but are not chemical binding [9, 10, 25–29]. To be regarded as the characteristic feature of a particular protein biological function, the RRM frequency should satisfy the following criteria:

1. single frequency/peak only can be observed for a group of protein sequences sharing the same biological activity;
2. if no significant or prominent peak in the spectrum can be seen, then these protein sequences are biologically unrelated; and
3. different peak frequencies observed in cross-spectrum correspond to different biological functions.

Once the RRM characteristic frequency for a particular protein function or interaction is determined, it is possible then to proceed with further calculations:

1. by using inverse Fourier transform (IFT), defined as:

$$x_n = \sum_{i=0}^{N-1} X_k e^{i*2\pi nk/N} \quad n \in Z$$

where $N$, number of frequencies analysed; $n$, amino acid considered; $x_n$, value of the signal at the position corresponding to the amino acid; $k$, current frequency we are considering (0 up to $N-1$); $X_k$, amount of frequency $k$ in the signal. By using inverse Fourier transform (IFT) it is possible to determine the so-called hot spot amino acids that contribute mostly to this frequency and, thus, to the observed function can be predicted/defined within the protein sequence.

2. by employing wavelet transform, the locations of a protein's active/binding sites can be predicted. In general, wavelet transformation of the signal can be described as:

$$\Psi_{a,b} = \frac{1}{\sqrt{|a|}}\Psi\left(\frac{t-b}{a}\right)$$

$\Psi_{a,b}$ is calculated using a scaling function $\Psi$ by time $b$ and scale $a$.

In RRM analysis, the Morlet wavelet function is used and defined as:

$$\omega(t) = C\mathrm{e}^{\left(\frac{-t^2}{2}+j\omega_0 t\right)}$$

3. de novo design short bioactive peptide analogues (on the basis of determined frequency and phase) expressing the same biological activity as the original protein sequence. The biological activities of computationally designed bioactive peptides were successfully evaluated in a number of experimental studies [1, 20, 40].

In a number of different protein examples it was shown that proteins and their targets (other proteins or small molecules) have the same characteristic frequency in their multiple cross-spectral functions defined as the magnitude of the normalized vector (signal-to-noise) product of RRM spectra in complex domain as follows [9]:

$$M = \left|M_1 * \bar{M}_2\right|$$

where $M_1$ and $M_2$ are RRM spectra of two proteins, $\bar{M}$ means complex conjugate.

Despite the fact that a protein and its target have different biological functions, they still can interact or participate in the same biological process, which is defined by the same RRM frequency [10, 25–29].

### 2.2 Time–frequency analysis

The Wigner function has been introduced to describe a phase space distribution for applications in both classical physics and quantum mechanics areas [44]. Thus, convolution of the Wigner functions of quantum state and filtered state can represent a phase space distribution of the analysed system [2, 21]. Moreover, the Wigner–Ville distribution (WVD) can be used to describe the changes in frequency content over a period of time. In the special case of a linear polymer (a protein sequence as an example), the WVD may represent the distribution of the energy of various frequency components of the signal at particular positions along the protein (distance between an amino acid is set at an arbitrary value $d = 1$).

The main problem in practical calculations, using convolution of the signal, is the cross-term that represents interference of the signals. To overcome loss of resolution due to cross-terms, we replaced the WVD by the smoothed pseudo Wigner–Ville distribution (SPWVD). Supposing EIIP[i], $i = 1, 2, \ldots, \mathrm{N}$ is the numerical sequence of the electron ion interaction potentials (EIIP) of amino acids along the polypeptide chain, then the SPWVD of the EIIP is given by [41]:

$$S(t,f) = \int_{-\infty}^{\infty} h(\tau)\int_{-\infty}^{\infty} g(s-t)z(s+\tau/2)z(s+\tau/2)^*\mathrm{d}s\,\mathrm{e}^{-j2\pi\nu\tau}\mathrm{d}\tau$$

where $z(s)$ is a complex signal (in our case a sequence of potentials) generated from the numerical sequence EIIP[i] by using a Hilbert transform; $z(s)^*$ is a complex conjugated from the signal; and $h(\tau)$ and $g(\tau)$ represent kernel functions (in our case windows function for frequency and time (space) smoothing); $t$ represents a time/spatial coordinate; $f$ is a frequency

In discrete form, the SPWVD can be calculated as:

$$W(n,m) = \frac{1}{2}N\sum_{k=-N+1}^{N+1}|h(k)|^2 \sum_{p=-M+1}^{M-1} g(p)z(n+p+k)z^*(n+p-k)\mathrm{e}^{-\frac{2i\pi km}{M}}$$

where $k$ is the unit in the frequency domain, $p$ is the point in the time/spatial domain, while $n$ and $m$ are the coordinates in $t$–$f$ plane corresponding to the position and frequency, respectively. The $h(k)$ and $g(p)$ represent independent frequency and time/spatial smoothing, respectively.

In this study, as the smoothing functions, we used the Gauss filters, which are defined as:

$$h(k) = \mathrm{e}^{(-k^2/2\sigma)/(\sigma\sqrt{2\pi})}; \quad g(p) = \mathrm{e}^{(-p^2/2\sigma)/(\sigma\sqrt{2\pi})}$$

$\sigma$ is the standard deviation and $k$ and $p$ are the mean frequency and distance. The resulting SPWVD is shown in a $t$–$f$ plane as a contour plot regarding the values of $S(t, f)$. Furthermore, the values of $S(t, f)$ are normalized by dividing calculated values with the average value calculated over all $t$–$f$ plane. It can be assumed that $S(t, f)$ represents the distribution of energy carried by a signal in the space domain [41]. In our previous studies, the SPWVD was

**Table 1** RRM frequencies, $f_{RRM}$, and normalized intensities, Nin*, calculated for each analysed HSP group (ordered by the normalized intensity value)

| Protein group | $f_{RRM}$ | Nin* | $f_{RRM}$ | Nin* | $f_{RRM}$ | Nin* |
|---|---|---|---|---|---|---|
| ID HSP | 0.081 | 34.6 | 0.267 | 13.5 | 0.42 | 9.5 |
| HSP27 | 0.285 | 18.2 | 0.267 | 14.5 | 0.489 | 13.5 |
| HSP60 | 0.103 | 20.4 | 0.065 | 13.2 | 0.341 | 12.9 |
| HSP70 | 0.169 | 14.9 | 0.354 | 13.3 | 0.267 | 5.4 |
| HSP90 | 0.267 | 46.0 | 0.366 | 18.0 | 0.080 | 9.5 |
| HSP (cancer related) | 0.355 | 29.0 | 0.065 | 26.5 | 0.267 | 6.5 |

incorporated in the RRM approach to predict the locations of active sites in selected proteins [41, 42].

In this study, we investigated the applicability of the SPWVD in prediction of ID regions in selected cancer-related HSPs as well as the locations of the active/binding sites within these proteins. SPWVD is not a new method in signal processing analysis. However, its application to protein structure–function analysis is novel. We found that the features of SPWVD are suitable for analysis of a signal derived from one-dimensional distribution of electron ion interaction potentials (EIIP) along a protein sequence. It is highly efficient for detection of the irregularities in charge density distribution along the protein [41]. Efficiency of the fast Fourier transformation (FFT) for $n$ points is defined as $n$*Log($n$). A number of calculations in the SPWVD depend on a number of points, $n$ (amino acids), in a given protein and are lower than $n^2$. Most proteins are less than 1000 amino acids in length; therefore, computational cost (processing time) is not an issue.

## 3 Results

### 3.1 Determination of RRM characteristic frequency for selected HSP proteins

In this study the standard RRM approach was used to determine RRM characteristic frequencies of HSPs that were grouped on the basis of their biological functions. ID mammalian HSP sequences were selected from the Database of Protein Disorder (DisProt http://www.disprot.org/).

The following sequences were used to determine the characteristic frequency of the ID HSP group (Table 1):

1. DisProt|DP00142|uniprot|P14602|unigene|Mm.13849|sp|HSPB1_MOUSE_2
2. DisProt|DP00358|uniprot|Q15185|unigene|Hs.50425|sp|TEBP_HUMAN_3
3. DisProt|DP00444|uniprot|P02489|unigene|Hs.184085|sp|CRYAA_HUMAN_alpha_crystaline_4
4. DisProt|DP00445|uniprot|P02511|sp|CRYAB_HUMAN_5
5. DisProt|HSPB8_human_Hsp22_13_6.

Small HSPs have low molecular masses (13–43 kDa) and contain a conservative α-crystallin domain (about 90 residues) that consists of several β-strands forming two β-sheets packed in immunoglobulin-like manner. The α-crystallin domain plays an important role in formation of stable small HSP dimers, which are the building blocks of the large HSP oligomers. The N-terminal domain and C-terminal extension are flexible and susceptible to proteolysis and post-translational modifications and are predominantly intrinsically disordered. These disordered N- and C-terminal sequences play important roles in the structure, regulation, and functioning of small HSP [35]. The RRM was applied here to analyse 5 mammalian ID HSPs (shown above), and their RRM characteristic frequency (most prominent) was identified at $f_{RRM} = 0.081$ (Table 1) and shown in Figs. 1a and 2a.

To determine the characteristic frequencies for cancer-related proteins from the HSP family (35 mammalian sequences), HSP27, HSP60, HSP70, and HSP90 protein sequences were selected from the UniProt database (Table 1). It was reported that several HSPs are involved with the prognosis of specific cancers. In particular, HSP27 are overexpressed in gastric, liver, and prostate carcinoma and osteosarcomas. HSP70 are found to be overexpressed in breast, endometrial, uterine cervical, and bladder carcinomas. HSP90 plays a particularly versatile role in cell regulation by forming complexes with a large number of cellular kinases, transcription factors, and other molecules. A role of HSP60 in cancer is uncertain; however, its upregulation or downregulation has been reported in various tumour series correlating with disease outcome [19].

According to the EMBL-EBI (http://www.ebi.ac.uk/interpro/protein/database), for each particular HSP group, the following domains were identified as functionally important:

HSP27—have 2 domains: HSP20-like chaperon and alpha-crystalline domain 76–83;
HSP60—have mostly chaperonin structure;
HSP70—have protein-binding domain 439–586 and heat-shock terminal 570–652;
HSP90—N-terminal 17–225 looks like a protein kinase, and 295–547 is ribosomal protein-like region.

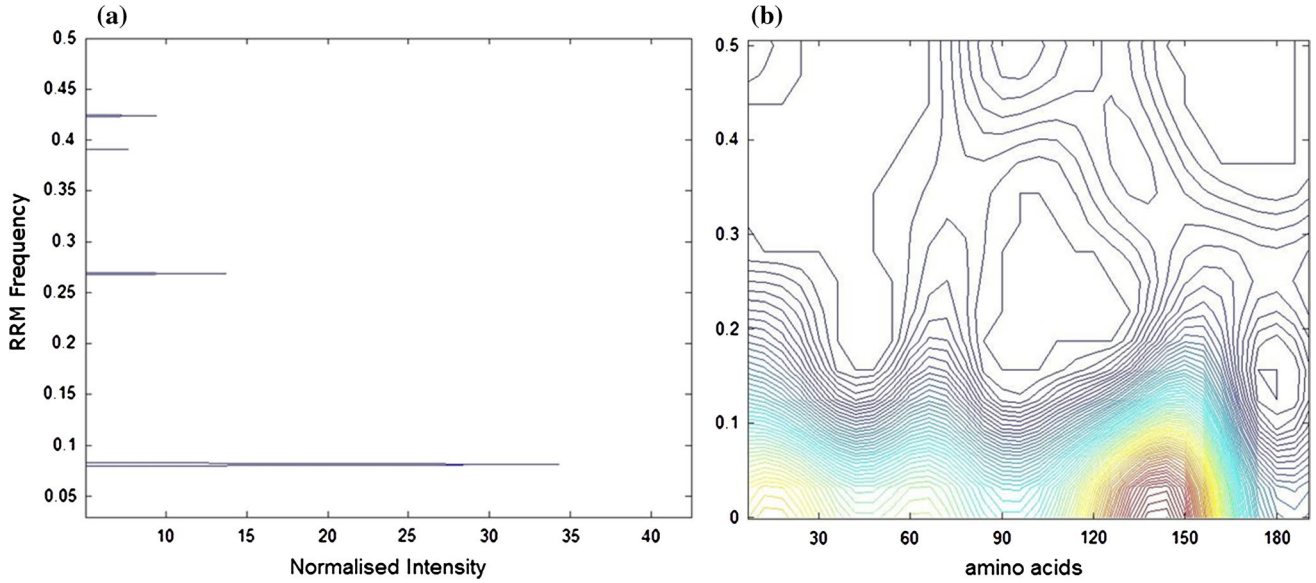



**Fig. 1** **a** Cross-spectral function of 5 mammalian ID HSP sequences; **b** *t–f* plane contour plot for human alpha-crystalline A chain P02489. Residue 138 is susceptible to oxidation, C-terminal extension: residues 140-175, determined by X-ray crystallography (UniProt data)

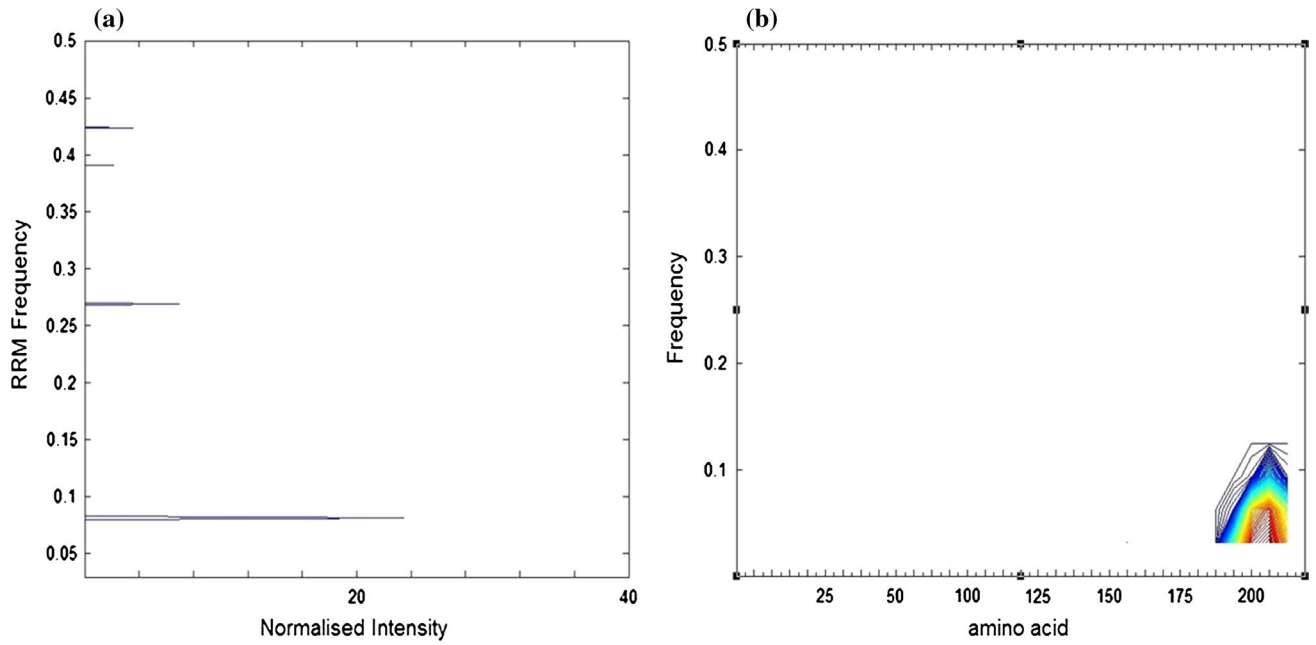



**Fig. 2** **a** Cross-spectral function of 5 mammalian ID HSP sequences; **b** *t–f* plane contour plot for P14602 HSPB1_MOUSE_2. Long-term orientation (LTO) ID region at 192–209, determined by X-ray crystallography (UniProt data)

For the RRM analysis, we formed the following functional groups of HSPs: HSP27 (7 sequences), HSP60 (5 sequences), HSP70 (13 sequences), HSP90 (10 sequences), and all HSP cancer related (35 sequences). Using the RRM, the characteristic frequency of HSP27 was identified at $f_{\mathrm{RRM}} = 0.285$ (Table 1; Fig. 3), for HSP60—at $f_{\mathrm{RRM}} = 0.103$ (Table 1; Fig. 4), for HSP70—at $f_{\mathrm{RRM}} = 0.169$ (Table 1; Fig. 5), and for HSP90—at $f_{\mathrm{RRM}} = 0.267$ (Table 1; Fig. 6). We also analysed the combined group of HSP27, HSP60, HSP70, and HSP90 (all HSP cancer related, 35 sequences), and the RRM frequency for these cancer-related HSPs was identified at $f_{\mathrm{RRM}} = 0.355$ (Table 1; Fig. 7). Interestingly, the frequency $f_{\mathrm{RRM}} = 0.267$ is common for HSP27, HSP70, HSP90, HSP cancer-related and ID HSP (Table 1).

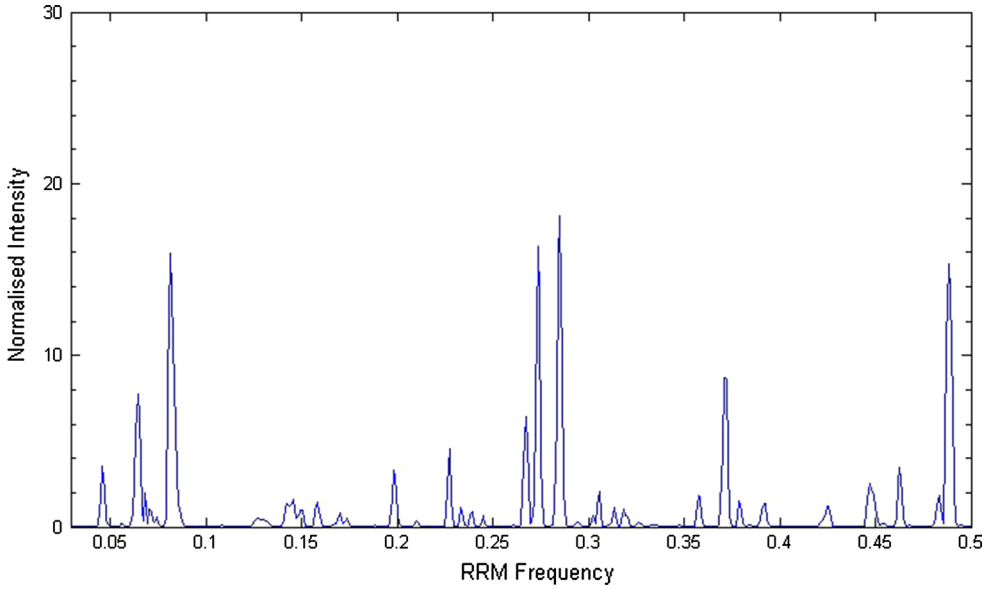



**Fig. 3** Cross-spectral function of 7 mammalian HSP27 sequences

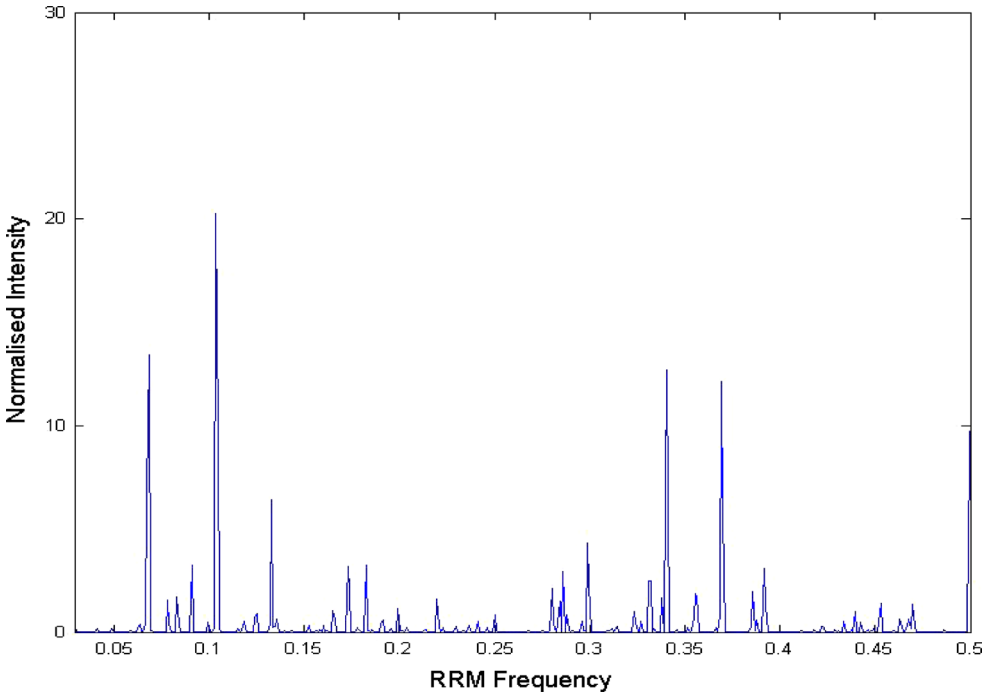



**Fig. 4** Cross-spectral function of 5 mammalian HSP60 sequences

This implies that all proteins within these groups have this frequency component (with different amplitude ratios) in common. According to the RRM theory, it means that they share the same biological activity. Each particular biological function or interaction is characterized by a unique RRM characteristic frequency [9, 10, 25–29]. From Figs. 3, 4, 5, 6, and 7 we can see that there are a number of prominent peaks in the cross-spectral functions of the corresponding protein groups. It reveals the multifunctional roles these proteins play, i.e. they can participate in different biological processes and/or interact with other partners.

### 3.2 Application of SPWVD for prediction of ID regions, active/binding sites in selected HSP sequences

The SPWVD transformation was used to identify locations of protein's active/binding sites on the basis of the determined RRM frequencies. Only four sequences with the ID regions were selected for analysis. These sequences were

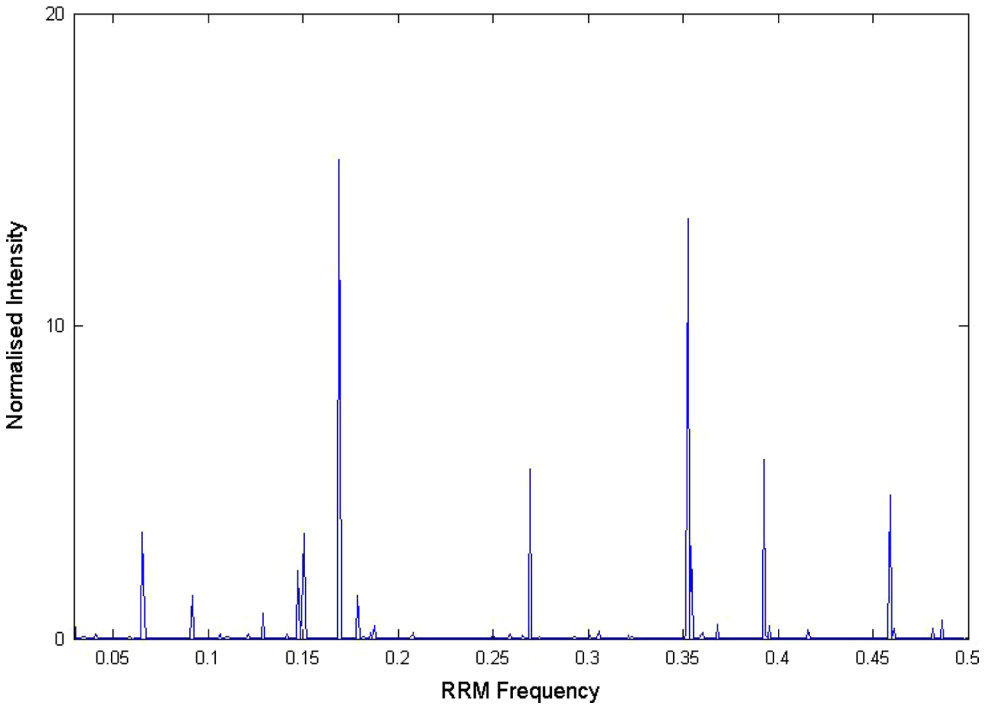



**Fig. 5** Cross-spectral function of 13 mammalian HSP70 sequences

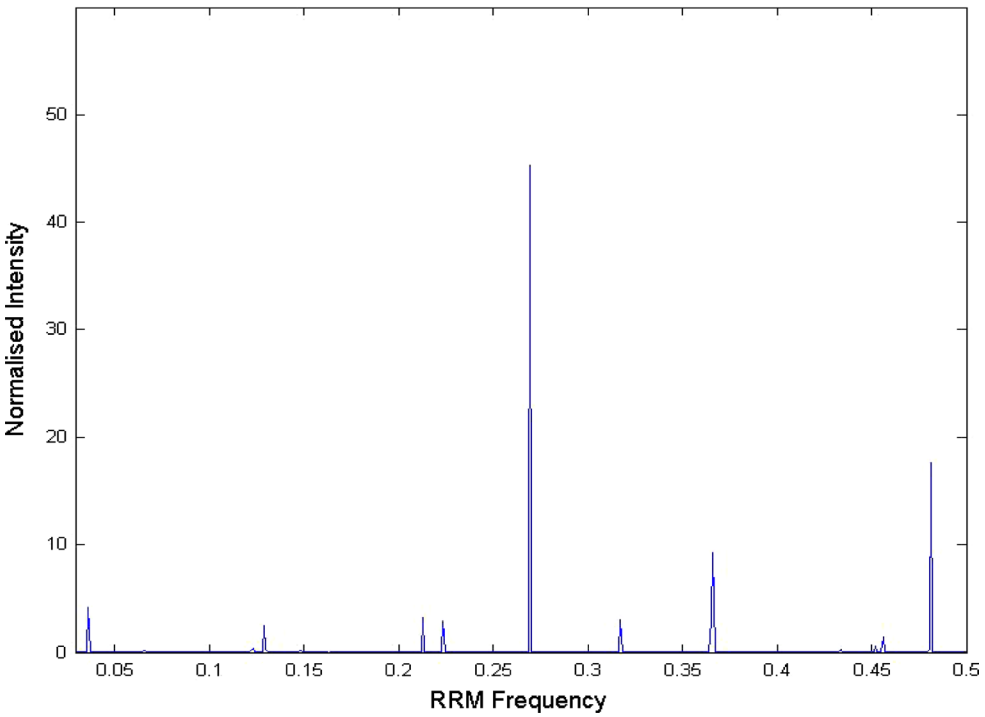



**Fig. 6** Cross-spectral function of 10 mammalian HSP90 sequences

analysed using SPWVD, PONDR, and MobiDB, with the results compared with the ID regions, determined experimentally by X-ray crystallography (where data are available), and regions predicted by the above-mentioned computational methods. SPWVD was applied to predict the location of active/binding sites for four selected ID HSP sequences. The results are presented in Table 2 and Figs. 1b, 2b, 8, and 9:

### *3.2.1 Human Alpha-crystalline A chain P02489*

P02489 HSP sequence was selected as an example for prediction of its disordered regions. Our calculations,

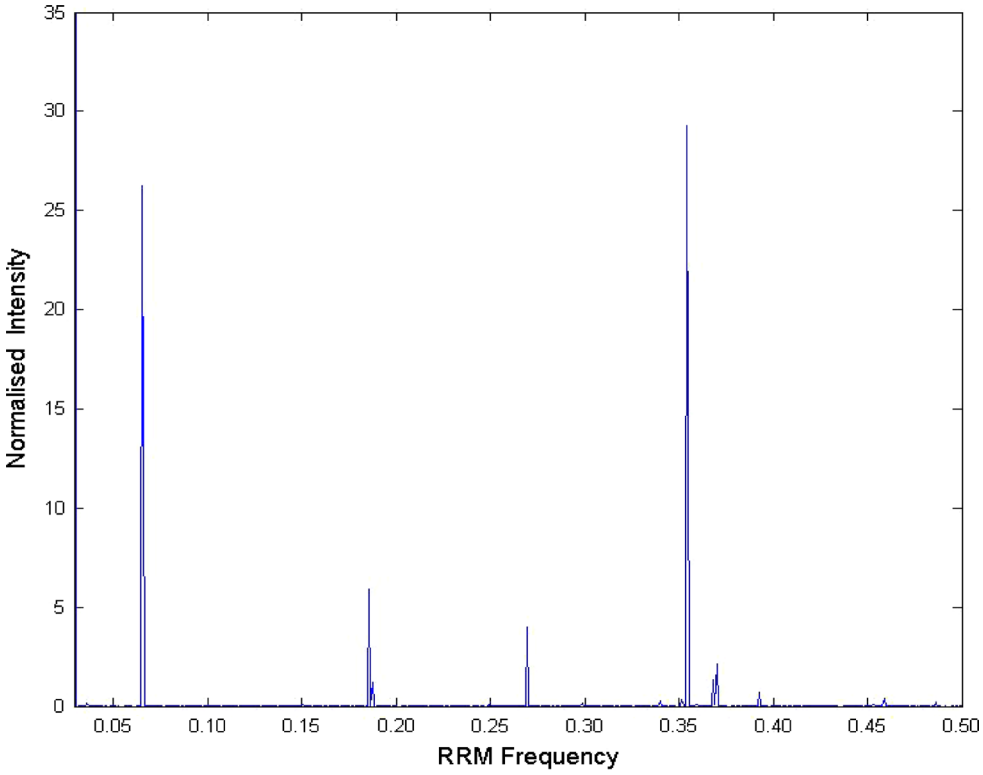



**Fig. 7** Cross-spectral function of the combined group (HSP27, HSP60, HSP70, and HSP90): 35 mammalian cancer-related sequences

**Table 2** ID regions predicted in selected HSPs: X-rays, RRM-SPWVD, MobiDB, and PONDR data

| Protein code | X-rays data | RRM-SPWVD | MobiDB | PONDR |
|---|---|---|---|---|
| P02489 | 140–175 | 125–175 | 144–173 | 1–3; 160–175 |
| P14602 | 192–209 | 187–209 | 188–209 | 140–209; 110–125 |
| P07900 | NA | 125–175; 380–407 | 163–184 + a few very short (2–3 aa) | 52–59;183–189; 225–300; 385–405 many very short (2–3 aa) |
| P0DMV8 | NA | 270–340 | NA | 270–290 and 310–340 |

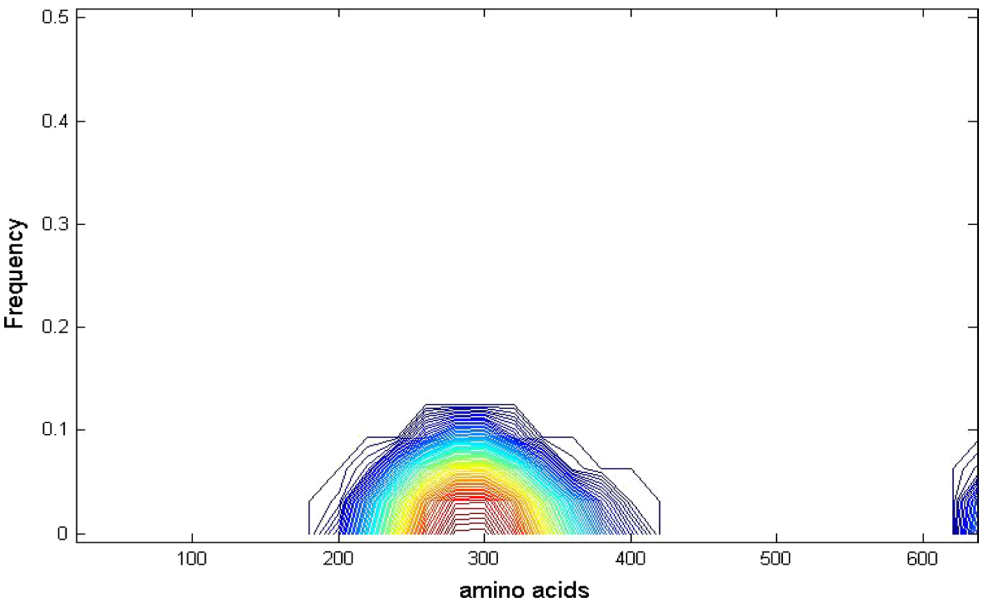



**Fig. 8** *t–f* plane as a contour plot long-term orientation (LTO) for P0DMV8 HSP_70_HS71A (UniProt data)

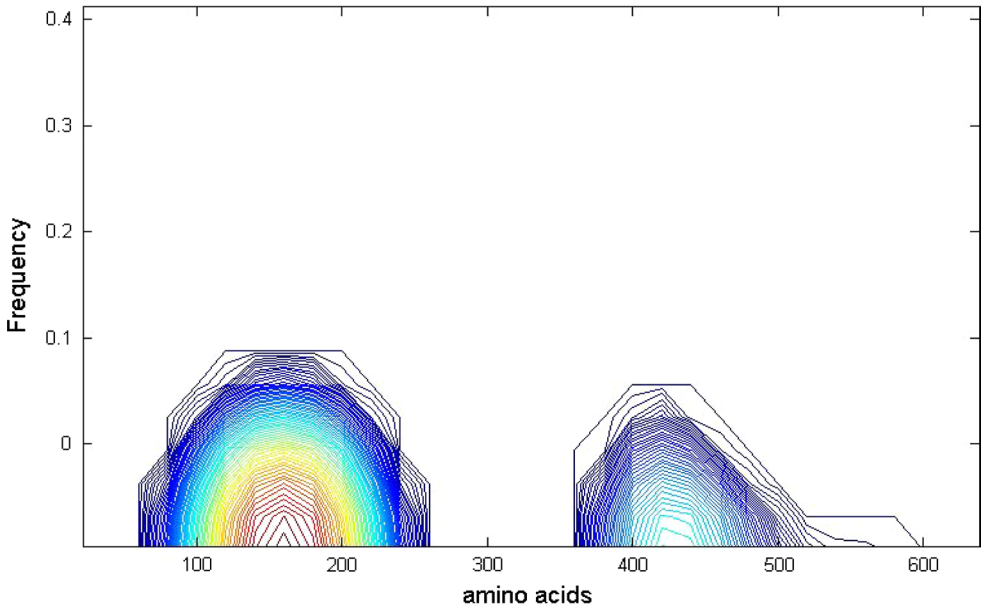



**Fig. 9** *t–f* plane as a contour plot long-term orientation (LTO) for P07900 HS90A_HUMAN (UniProt data)

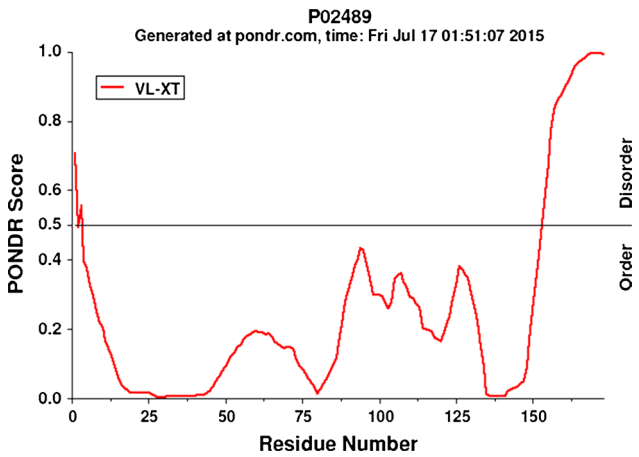



**Fig. 10** PONDR score for P02489 protein sequence was generated by PONDR tool

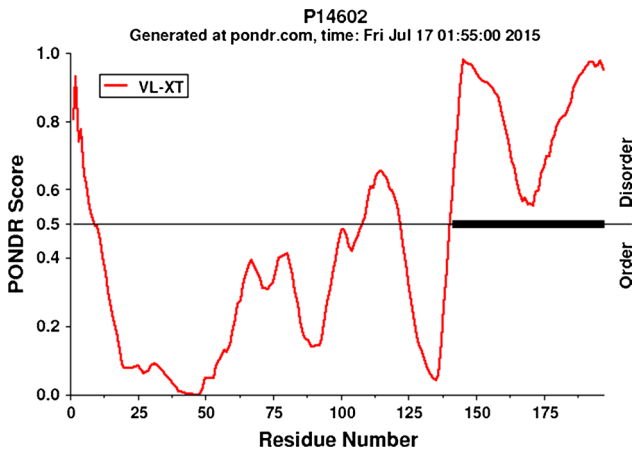



**Fig. 11** PONDR score for P14602. The figure was generated using PONDR tool

performed using the SPWVD, showed that the amino acids mostly contributing to the ID HSPs characteristic frequency $f_{\mathrm{RRM}} = 0.081$ (Table 1; Fig. 1a) are located at 125–175, and this region covers the functionally important residue 138 and partially covers the intrinsically disordered C-terminal extension (Table 2; Fig. 1b). The experimentally determined disordered region is located at 140–175 (X-ray data). The analysis of P02489 protein sequence, using the MobiDB 2.0 tool, reveals that the disordered region is located at 144–173. In addition, a few, short disordered regions were also identified using this predictor (please refer to Supplementary material, Appendix I). The application of the PONDR tool predicted two segments, 160–175 and 1–3 (Fig. 10), as the positions of disordered regions in P02489 protein (Table 2).

### *3.2.2 P14602 hspb1_mouse_2*

We also applied SPWVD transformation to analysis of another protein, P14602. The results reveal that the amino acids mostly contributing to the ID HSPs characteristic frequency $f_{\mathrm{RRM}} = 0.081$ are located at 187–209, and this location covers the functionally important ID region 192–209, which was determined experimentally by X-ray crystallography (Table 2; Fig. 2b). For P14602 protein, the predicted positions of the disordered regions vary significantly depending on the specific predictor used in the analysis. For example, MobiDB consensus tool predicted several short segments and one longer segment of the disordered properties located close to the C end of the protein, 188–209 (Table 2; Supplementary material, Appendix II). By

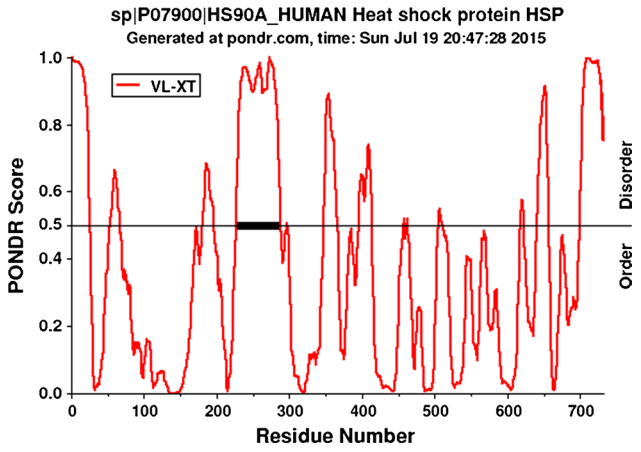



**Fig. 12** PONDR score for P07900. The figure was generated using PONDR tool

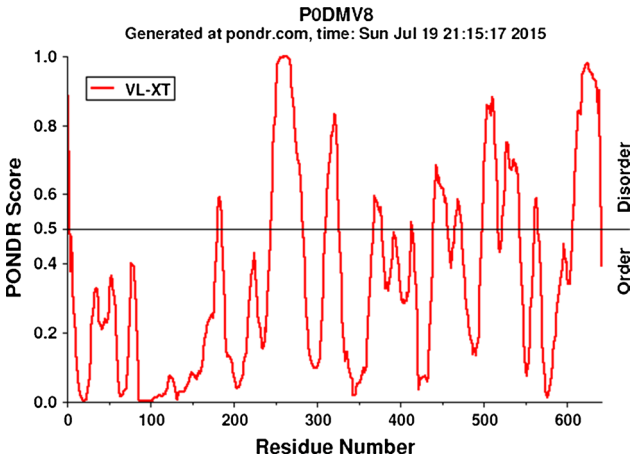



**Fig. 13** PONDR score for P0DMV8. The figure was generated using the PONDR tool

using the RRM-SPWVD method, we achieved very similar results with the predicted disordered region to be at 187–209, while the X-ray experimental data (www.disprot.org) determined the region of 192–209 to be disordered. Interestingly, the PONDR tool revealed the segments 140–209 and 110–125 as the disordered regions in the P14602 protein sequence (Table 2; Fig. 11).

#### *3.2.3 P07900 heat-shock protein HSP 90-alpha*

P07900 protein is an important molecular chaperone. The RRM-SPWVD method predicted the location of two segments to be disordered, namely 125–175 and 380–470 (Table 2; Fig. 9). MobiDB 2.0 tool predicted the disordered region to be at 163–184, with a few additional short segments (Table 2; Supplementary material, Appendix III). Application of the PONDR tool resulted in the prediction of a number of disordered regions, with the main ID region positioned at 225–300. There are also two short regions, predicted by PONDR, that are located at 52–59 and 183–189 (Fig. 12). These predicted regions do not overlap with the locations predicted by the RRM-SPWVD. However, the last region predicted by PONDR and located at 385–400 is overlapping with the larger region 380–407 predicted by the RRM-SPWVD (Table 2; Fig. 9).

#### *3.2.4 P0DMV8 heat-shock 70 kDa protein 1A*

Using the RRM-SPWVD, we predicted the disordered regions in P0DMV8 protein located at 270–340 (Table 2; Fig. 8). Application of the PONDR tool resulted in the predicted disordered regions positioned at 270–290 and 310–340 (Table 2; Fig. 13). We cannot provide data for analysis of P0DMV8 protein using MobiDB 2.0 prediction tool as this particular protein is not available in the database. X-ray data are also unavailable.

## 4 Discussion

The RRM was used in this study to analyse the existence of patterns specific for the selected HSP proteins. Using the model, the characteristic frequencies for the ID and cancer-related HSPs were determined. The RRM is based on Fourier transform. The main disadvantage of the Fourier transform is that the information about frequency characteristic along the series is hidden, and only averaged time and frequency content of the analysed signal can be obtained. As was mentioned above, on the basis of the RRM characteristic frequency determined for the particular biological function or interaction, it becomes possible to identify the individual "hot spot" amino acids that contributed most to this specific characteristic frequency and thus, possibly to the observed biological behaviour of the protein. This can be achieved by using the IFT. Our previous studies showed that the identified "hot spot" amino acids are clustered in and around a protein's active sites.

In the last 20 years, the time–frequency distribution methods have become powerful alternative tools for signal analysis. A time–frequency transform presents energy distribution of a signal over the time and frequency domains. In this study, we applied the time–frequency signal processing technique to the selected HSPs, aiming to demonstrate how a signal's energy is distributed over the time–frequency space. By incorporating smoothed pseudo Wigner–Ville distribution (SPWV) in the standard RRM approach, we overcome the problem of non-localized events currently present in the model. In particular, we have shown that by knowing a protein's specific RRM characteristic frequency and applying the SPWVD to an individual protein sequence, it becomes possible to predict the position of this protein's active sites (functional epitopes) in the protein. Here, we compared the predictive capability of the

RRM-SPWDV and the computational tools, MobiDB2.0 and PONDR, in analysis of ID regions and protein active/binding site of the HSPs.

In comparison with the predictors, MobiDB2.0 and PONDR, our method, RRM-SPWVD, is based on analysis of charge distribution along the whole polypeptide chain (as opposed to the short windows of amino acids regions) characterized or presented by the EIIP of each amino acids. The important feature of our approach is that it is not using neural network training, which is based on similarity between the studied protein of interest and known intrinsically disordered proteins. Instead, the RRM-SPWVD approach allows determining disordered regions within a protein sequence by analysing the information written in its whole primary structure. To retrieve information about disruption in periodicity in charge density along the protein sequence, that can match ID regions along the protein molecule, we use space frequency analysis.

The results obtained show that our computational predictions using the RRM-SPWVD method correspond closely with the experimentally identified locations (where available) and are in agreement with the locations of the ID regions determined by the MobiDB 2.0 and PONDR tools (Table 2). In particular, the RRM-SPWVD-predicted ID regions are similar to the longer regions identified by these standard predictors. The findings also revealed that our method could not identify very short regions of ID (1–2 amino acids in length) when compared to the results obtained by MobiDB 2.0 and PONDR computational tools (Table 2). On the other hand, prediction of very short ID regions by these two tools shows a high degree of variability in results and significant differences in predicted locations, when compared with the experimental X-ray data (please refer to Figs. 10, 11, 12, 13, and Appendix 1).

In essence, the incorporation of the SPWVD in the RRM: (1) enables not only prediction of the functionally important amino acids (the so-called point mutations as done in the standard RRM using the inverse Fourier transform (IFT) [9, 25]) but prediction of the regions of functional importance, such as active/binding sites and ID regions along the analysed HSP sequences; (2) allows reducing a number of proteins required for efficient analysis in order to get an accurate computational prediction. In particular, we can calculate the RRM frequency by using a limited number of protein sequences (from one to three proteins sequences).

## 5 Conclusion

Our previous computational studies, employing the RRM approach, demonstrated that digital signal processing methods can be applied successfully to analysis of the informational content of different protein primary sequences. Fourier transform has been used in the RRM to determine a protein's characteristic frequency corresponding to its particular biological activity. However, in protein structure–function analysis studies, it is of particular importance to be able to predict accurately the locations of active and/or binding sites within these proteins. Due to limitations of the classical non-localized spectral transformations, we applied here the SPWVD instead of the one-dimensional Fourier transform previously used within the RRM. Up to date this new tool has been tested on the selected proteins such as cytochrome C, glucagon, haemoglobin, and oncogene protein Ha-ras p21 [41, 42]. In this study, the SPWVD, incorporated into the RRM, was used to predict the locations of intrinsic and binding sites in the selected cancer-related HSPs. The findings clearly show that substitution of the Fourier transform with the SPWVD within the RRM improves the prediction of protein active/binding sites allocation, as opposed to prediction of individual "hot spot" which found to be clustered in and around the active site. Incorporation of the SPWVD in the RRM enabled to address its limitation. The SPWVD spectrum allows presenting the identified disordered region as a *t–f* plane (the region is presented in the frequency and time/space domains). The prediction accuracy depends on the size of a given protein, and it is defined in terms of a resolution of spectra, $1/N$.

The selected HSP sequences were also analysed using the standard predictors, MobiDB 2.0 and PONDR, with the results being compared with the predictions obtained by the RRM-SPWVD method. Our predictions are in agreement with the locations of the ID regions determined by MobiDB 2.0 and PONDR. Our results correspond closely with the experimental data obtained by X-ray crystallography. The findings of this computational study suggest that the proposed RRM-SPWVD method based on signal processing techniques can be used for accurate predicting the locations of the ID regions and active/binding sites in the selected HSP sequences.

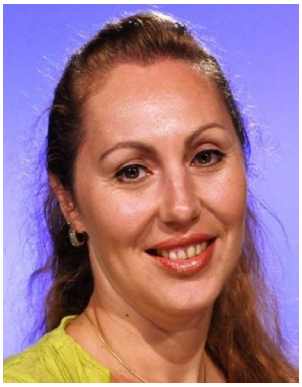
## References

1. Almansour NM, Pirogova E, Coloe PJ, Cosic I, Istivan TS (2012) Investigation of cytotoxicity of negative control peptides versus bioactive peptides on skin cancer and normal cells: a comparative study. Future Med Chem 4(12):1553–1565
2. Boashash B (2005) Time–frequency signal analysis and processing. Prentice Hall, Upper Saddle River
3. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol 14:292–299. doi:10.1016/j.sbi.2004.05.003

4. Calderwood SK, Stevenson MA, Murshid A (2012) Heat shock proteins, autoimmunity, and cancer treatment. Autoimmune Dis. doi:10.1155/2012/486069
5. Chuma M, Sakamoto M, Yamazaki K, Ohta T, Ohki M, Asaka M, Hirohashi S (2003) Expression profiling in multistage hepatocarcinogenesis: identification of HSP70 as a molecular marker of early hepatocellular carcinoma. Hepatology 37:198–207
6. Ciocca DR, Jorge AD, Jorge O et al (1991) Estrogen receptors, progesterone receptors and heat shock 27-kD protein in liver biopsy specimens from patients with hepatitis B virus infection. Hepatology 13:838–844
7. Ciocca DR, Clark GM, Tandon AK, Fuqua SAW, Welch WJ, McGuire WL (1993) Heat shock protein Hsp70 in patients with axillary lymph node-negative breast cancer: prognostic implications. J Natl Cancer Inst 85:570–574
8. Ciocca DR, Green S, Elledge RM et al (1998) Heat shock proteins Hsp27 and Hsp70: lack of correlation with response to tamoxifen and clinical course of disease in estrogen receptor-positive metastatic breast cancer (a Southwest Oncology Group study). Clin Cancer Res 5:1263–1266
9. Cosic I (1994) Macromolecular bioactivity: is it resonant interaction between macromolecules? Theory and applications. IEEE Trans Biomed Eng 41:1101–1114
10. Cosic I, Lazar K, Cosic D (2014) Prediction of tubulin resonant frequencies using the resonant recognition model (RRM). IEEE Trans NanoBiosc. doi:10.1109/TNB.2014.2365851
11. Cruz-Barbosa R, Vellido A, Giraldo J (2014) The influence of alignment-free sequence representations on the semi-supervised classification of class CG protein-coupled receptors. Med Biol Eng Comput 53(2):137–149
12. Deng X, Eickholt J, Cheng J (2012) A comprehensive overview of computational protein disorder prediction methods. Mol BioSyst 8:114–121
13. Domenico TD, Walsh I, Martin AJM, Tosatto SCE (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 28(15):2080–2081
14. Dosztanyi Z, Meszaros B, Simon I (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform 11:225–243
15. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41:6573–6582
16. Fong JH, Shoemaker BA, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV, Panchenko AR (2009) Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. PLoS Comput Biol 5(3):1–11
17. Hafeez A, Asghar W, Rafique MM, Iqbal SM, Butt AR (2012) GPU-based real-time detection and analysis of biological targets using solid-state nanopores. Med Biol Eng Comput 50(6):605–615
18. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19:929–949
19. Hjerpe E, Egyhazi S, Carlson J, Stolt MF, Schedvins K, Johansson H, Shoshan M, Avall-Lundqvist E (2013) HSP60 predicts survival in advanced serous ovarian cancer. Int J Gynecol Cancer 23(3):448–455
20. Istivan TS, Pirogova E, Gan E, Almansour NM, Coloe PJ, Cosic I (2011) biological effects of a de novo designed myxoma virus peptide analogue: evaluation of cytotoxicity on tumor cells. PLoS ONE. doi:10.1371/journal.pone.0024809
21. Lalovic D, Davidovic DM, Bijedic N (2003) Quantum mechanics in terms of non-negative smoothed Wigner functions. Phys Rev A 46:1206–1212
22. Metallo SJ (2010) Intrinsically disordered proteins are potential drug targets. Curr Opin Chem Biol 14(4):481–488
23. Monastyrsky B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K (2014) Assessment of protein disorder region predictions in CASP10. Proteins 82(Suppl. 2):127–137
24. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK (2008) Intrinsic disorder in protein-protein interaction networks: case studies of complexes involving p53 and 14-3-3. BMC Genom 9(Suppl 1):S1. doi:10.1186/1471-2164-9-S1-S1
25. Pirogova E, Fang Q, Akay M, Cosic I (2002) Investigation of the structure and function relationships of oncogene proteins. Proc IEEE 90(12):1859–1867
26. Pirogova E, Akay A, Cosic I (2007) Computational analysis of interactions between tumor and tumor suppressor proteins. In: Akay M (ed) Genomics and proteomics engineering in medicine and biology. Wiley, Hoboken, pp 257–287
27. Pirogova E, Akay M, Cosic I (2009) Investigating the interaction between oncogene and tumor suppressor protein. IEEE Trans IT Biomed 13(1):10–15
28. Pirogova E, Vojisavljevic V, Caceres J, Cosic I (2010) Ataxin active site determination using spectral distribution of electron ion interaction potentials of amino acids. Med Biol Eng Comput 48(4):303–309
29. Pirogova E, Istivan T, Gan E, Cosic I (2011) Advances in methods for therapeutic peptide discovery, design and development. Curr Pharm Biotechnol 12(8):1117–1127
30. Potenza E, Di Domenico T, Walsh I, Tosatto SCE (2014) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucl Acids Res. doi:10.1093/nar/gku982
31. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK (2007) Intrinsic disorder and functional proteomics. Biophys J 92:1439–1456
32. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered protein. Proteins 42(1):38–48
33. Russell RB, Gibson TJ (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies. FEBS Lett 582(8):1271–1275. doi:10.1016/j.febslet.2008.02.027
34. Spolar RS, Record MT Jr (1994) Coupling of local folding to site-specific binding of proteins to DNA. Science 263:777–784
35. Sudnitsyna MV, Mymrikov EV, Seit-Nebi AS, Gusev NB (2015) The role of intrinsically disordered regions in the structure and functioning of small heat shock proteins. Curr Protein Peptide Sci. doi:10.2174/138920312799277875
36. Tompa P, Csermely P (2004) The role of structural disorder in the function of RNA and protein chaperones. FASEB J 18:1169–1175
37. Uversky VN (2013) A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci 22:693–724
38. Uversky VN, Dunker AK (2010) Understanding protein non-folding. Biochim Biophys Acta 1804:1231–1264
39. Veljkovic V, Slavic I (1972) Simple general-model pseudopotential. Phys Rev Lett 29:105
40. Veljkovic V, Veljkovic N, Aüther J, Dietrich U (2007) Application of the EIIP/ISM bioinformatics concept in development of new drugs. Curr Med Chem 14:441–453
41. Vojisavljevic V, Pirogova E, Davidovic D, Cosic I (2009) A new approach to revealing functional residues from analysis of protein primary structure. In: Zhi-Pei L (ed) Proceedings of the 31st annual international conference of the ieee engineering in medicine and biology society. Institute of Electrical and Electronics Engineers, IEEE Xplore, pp 4731–4734
42. Vojisavljevic V, Pirogova E, Cosic I (2009) Studying repressors-DNA interactions using the RRM approach and time-frequency analysis. In: Proceedings of 2009 international symposium on bioelectronics & bioinformatics (ed), IEEE, USA, IEEE Xplore, pp 53–56

43. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN (2007) Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. J Proteome Res 6:1899–1916
44. Wigner EP (1932) On the quantum correction for thermodynamic equilibrium. Phys Rev 40:749–751
45. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res 6:1882–1898
46. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN (2007) Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J Proteome Res 6:1917–1932

**Dr. Vuk Vojisavljevic** completed B.E. in Electrical Eng and Physics, M.Sc. in Biophysics and Molecular Biology, University of Belgrade, Serbia, and Ph.D. in Biomed Eng, RMIT University, Australia. He is a Research Fellow at RMIT University.



**Associate Professor Elena Pirogova** completed B.E. (Hons) in Chemical Eng, National Technical University of Ukraine, Ukraine, and Ph.D. in Biomed Eng, Monash University, Australia. She is a Program Manager of Biomedical Engineering Degree at RMIT University.